B&B: Planning Bus Routes with Sharing-bikes in the City

Yue Yu* Tong Xia* Yong Li*

Abstract

Recently, the emergence of sharing-bikes exerts a significant impact on citizens' daily travel. However, when planning bus routes, the current system still does not take them into consideration. In this paper, we propose B&B, a data-driven system to plan bus routes with the consideration of sharing-bikes. Aiming to maximize the travel flow coverage, we design a heuristic approach to extend each bus route. During the route selection process, we propose a function to determine whether a road segment should be designed as a bus route or it is more suitable for passengers to ride sharing-bikes in that area. Besides, to ensure the efficiency of our proposed bus routes, a constraint is set up on the directness of bus routes. Extensive evaluation of two large-scale datasets in New York City demonstrates that our system achieves the best coverage on both flows and areas, which outperform by 7.05% and 15.55% over the baselines which do not consider sharing-bikes. Our system lays a solid foundation for urban planners to plan the city transportation systems overall, especially concerning the design of the bus routes and sharing-bike lanes.

1 Introduction

With the development of the city and the advocacy of green-life, modern citizens are more willing to choose a greener lifestyle. Moreover, sharing-bikes are available everywhere in many cities and countries, especially in China [19]. Their flexibility and convenience have significantly influenced the way people taking public vehicles. Citizens are more likely to ride a sharing-bike to take the public vehicles including metro and buses. Due to the high cost of construction of the metro, it is difficult to change its operation routes. Luckily, the bus routes are much more easy to adjust. Therefore, bus route planning system with the consideration of sharing-bikes to provide more efficient service is urgently needed.

Previous work has studied the problems of bus route planning [6, 7, 12] and bike lane design by data-driven methods [4]. However, they are investigated separately, and no work planned bus routes with the sharing-bikes. In this paper, we intend to explore the bus route planning problem with sharing-bikes to meet people's requests of travel and green-life.

Despite of its great significance, it is non-trivial to plan bus route with the consideration of sharing-bikes due to several challenges: *1*) How to design the bus routes in the city to meet the majority of travel demand while facilitating the bike-riding to bus station. If we design the bus routes densely, it is more convenient for passengers to take. However, under the limited bus route length, the coverage of overall bus service would be small and the travel efficiency would decrease. *2*) How to determine whether to plan a road as bus route when sharing-bike is available. To reduce the travel time and serve more traffic flow, it is reasonable to design more bus routes but the efficiency of the bus system would be discounted, as riding a bike is more preferred in short trips. This is to say, we need to trade off the chose of bus and bike. *3*) How to make all the routes as direct as possible to improve the user's travel experience. If a route is zigzagging, it can meet more travel demand, but bus passengers would not satisfied, especially for long-distance passengers. Therefore, the trade-off between meeting more travel demand and providing satisfactory service is also challenging.

To overcome these challenges, we propose a data-driven approach, named B&B, to plan **B**us routes with sharing-**B**ikes. For the first challenge, we detect some intensive pick-up and drop-off locations all over the city as bus station candidates and extend the bus routes among them to serve the majority of travel demand. While for travel between the near places, it is more convenient for passengers to ride sharing-bikes. By doing so, requests for both long and short distance trips can be satisfied at the same time. For the second challenge, we define a decision function to determine when to plan a road segment as the bus route based on the patterns of sharing-bikes trips, which is utilized in the dynamic expansion process for each bus route, as short-distance while small-flow lanes would be filtered for passengers to ride bikes, while lanes with longer distance or larger flow would be selected as bus routes. To solve the third challenge, we define a directness bound for bus route to maximize the flow covered by the route network with limited number and length of bus routes as well as restricted directness. Our contributions can be summarized as follows:

- We investigate the problem of planning bus routes with sharing-bikes. To the best of our knowledge, this is the first study that attempts to design the

*Tsinghua National Laboratory for Information Science and Technology. Department of Electronic Engineering, Tsinghua University, Beijing, China. liyong07@tsinghua.edu.cn.

bus route with the consideration of sharing-bikes by a data-driven method.

- We formulate the bus route planning problem as an optimization problem with three practical constraints, which maximize the coverage of passengers' travel demand all over the city with the sharing-bikes. We propose a three-stage heuristic approach to solve the formulated problem, which make as trade off the choose of bus and bike. Moreover, three kinds of initialization methods guarantee its effectiveness under different parameters.
- We evaluate our system using two large-scale datasets of New York City. Compared with the state-of-the-art baselines, our method achieves the best performance across different parameters. Specifically, it outperforms the benchmark by 7.05% and 15.55% in terms of traffic flow and area coverage without considering sharing-bikes.

The rest of this paper is organized as follows. To begin with, we systematically review the related works in Section 2. In section 3, we formulate the problem and provide an overview of our system. Motivated by the challenges, we introduce our method including preprocessing and route design in Section 4. After that, we apply our system to two real-world large-scale datasets in New York City and conduct an extensive analysis of the derived results to demonstrate that our methods have a better performance than baselines in Section 5. Finally, we conclude our paper in Section 6.

2 RELATED WORK

Data-Driven Urban Planning: With the availability of large-scale mobility data from smartphones, vehicles and transport systems, data-driven urban traffic planning techniques become increasingly popular [23]. Yuan et al. [21] demonstrated the existence of different functions regions in a city through GPS trajectory datasets. Wang et al. [16] built a network representation of human movement based on vehicle GPS tracks and extracted relevant clusters. With the popularity of the sharing-bike system, researchers investigated the problem of improving the public transportation system. For example, Nair et al. [14] analyzed bicycle data from Paris and uncovered the relationship between bicycle usage and multimodal trips, thus providing critical insights into station placement policy. In this paper, we focus on providing a data-driven approach to design the bus route of a city using large-scale OD data.

Taxi Trajectory Mining: A significant amount of literature aim to mine the trajectories of taxicabs since the trajectory data has recently become widely available. Related works about taxi recommendation [9], taxi ride-sharing service [13] and taxi demand prediction [20] have been published. Wei et al. [11] utilized a very large volume of real taxi trajectories in Beijing to detect outliers in spatio-temporal traffic. Zheng et al. [24] used the GPS trajectories of taxis in urban areas to evaluate the effectiveness of the carried out planning. Since taxi trajectories reflect citizens' travel demand, we can utilize the origin and destination of each taxi trajectory to design bus routes.

Bus Route Designing: Urban transportation network design problem includes two parts, one is the road network design problem, and the other is the public transit network design problem [8]. In this paper, we mainly focus on the latter problem. To solve this, the traditional method is to optimize an objective function with several constraints. For example, to maximize the area covered with the limited number of runs and frequency bounds for each line [22], heuristic methods [6], genetic algorithms [15] and other mathematical approaches [5] have been used to solve the optimization problem. Different from those works, we utilize heuristic while dynamic method to design bus route, and most importantly, we take the sharing-bike into consideration during the route expansion process.

3 Overview

In this section, we first attempt to model and define the problem of planning bus routes with the consideration of sharing-bikes, and then we introduce the framework of our solution.

3.1 Problem Formulation.

First, we introduce several concepts for the better explanation.

Definition 1: Origin-Destination (OD). The OD represents passengers' origin or destination of each travel. In this paper, we uniformly call the up place as origin and down place as destination, respectively.

Definition 2: Hot-spot. The hot-spot represents the location with highly intensive ODs. To effectively plan the bus routes, we assume the bus stations should be selected from those hot-spots.

Given the identified hot-spots, we build a weighted graph $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$ to simulate travel demand, where the vertex set $\boldsymbol{V}$ denotes the station candidates derived from hot-spots and the distance matrix $\boldsymbol{D}$ denotes the distance among all candidate stations. For a pair of nodes $v_i, v_j \in \boldsymbol{V}$, we say there exists an edge between v_i and v_j in $\boldsymbol{E}$, if there is trajectory records between v_i and v_j, and the flow matrix $\boldsymbol{F}$ represents the intensity of such traffic flow. For readability, we summarize the major notations used throughout the paper in Table 1.

Based on the flow network, our system aims to optimize the total flow served by both buses and sharing-bikes denoted by f under several practical constraints. Here, we define the set of bus routes as

Table 1: List of commonly used notations.

Notat.	Description
$\boldsymbol{V}$	The set of all candidate bus stations.
$\boldsymbol{F}$	The flow matrix for all candidate stations.
$\boldsymbol{D}$	The distance matrix for all candidate stations.
$\boldsymbol{L}$	The set of all the designed bus routes.
$\boldsymbol{B}$	The set of all trips fulfilled via sharing-bikes.
$\boldsymbol{C}$	The connectivity matrix among all candidate stations.
$r(v_m^i, v_n^i)$	The directness between station m and n in route i.
B_m	The length budget for every bus routes.
l_m	The maximum length for sharing-bike trips.
N	The maximum number of bus routes.

$\boldsymbol{L} = \{L_1, L_2, ..., L_i, ...\}$, where the total number of the bus routes is denoted as N_L. The *i-th* route is denoted as $L_i = (v_1^i, v_2^i, ..., v_m^i)$, where m is the number of stops in L_i, i.e. s_i, and $v_j^i \in \boldsymbol{V}(j = 1, 2, ..., m)$ stands for the *j-th* station in route i. Below are some general criteria for the planning of the route:

Criterion 1: The operation as well as the management cost for the bus system will increase rapidly with the growing number of routes. Thus, there should be an upper bound N for the number of bus routes.

Criterion 2: There should be an upper bound on the length of each route identically since the volume of fuel in each bus is limited. Moreover, if the route is too long, then the operation order of the bus lines are susceptible to traffic jam and traffic accidents. The maximum length of each route is set to be B_m.

Criterion 3: To make the transit operation of buses more efficiently, there should be restrictions on bus routes to prevent them from taking zigzag routes. We use the ratio $r(v_m^i, v_n^i)$ to measure the directness between station m and n in route i, which can be calculated as follows,

$$r(v_m^i, v_n^i) = \frac{\sum_{j=m}^{n-1} \boldsymbol{D}_{v_j^i, v_{j+1}^i}}{D_{v_m^i, v_n^i}} \quad (m < n). \tag{3.1}$$

The directness ratio should not be too large as the deviation of the route will make it lose passengers due to the caused relatively long commute time.

Optimization Goal. The ultimate goal is to design optimal bus routes to maximize the overall coverage of the traffic flow with the sharing-bikes.Based on the criteria above, we define the total flow f as the following: Given a set of bus routes, we construct a connectivity matrix $\boldsymbol{C}$ with

$$\boldsymbol{C}_{ij} = \begin{cases} 1 & travel\ between\ i, j\ is\ satisfied \\ 0 & otherwise \end{cases} \tag{3.2}$$

to document whether two vertices in our route network are connected. In this case, we assume that the passengers are willing to use the service of bus and sharing-bikes when their travel origin v_o and destination v_d are connected by buses and sharing-bikes (i.e. $\boldsymbol{C}_{v_o, v_d} = 1$), and the route are direct enough (i.e. $r(v_o, v_d) \leq \alpha$). With the definition, the optimization problem can be expressed as below:

$$\begin{aligned} \max \quad & \sum_{i<j, \boldsymbol{C}_{ij}=1, r(i,j)\leq\alpha} (\boldsymbol{F}_{ij} + \boldsymbol{F}_{ji}), \\ s.t. \quad & N_L \leq N, \\ & \sum_{j=1}^{s_i - 1} \boldsymbol{D}_{v_j^i, v_{j+1}^i} \leq B_m, \\ & r(v_m^i, v_n^i) \leq \alpha, \ \forall i \quad (1 \leq m < n \leq s_i). \end{aligned} \tag{3.3}$$

Our goal is to find a set of bus routes $\boldsymbol{L}$ to maximize the total flow in Equation 3.3 under the constraints 1-3. The first two criteria are aiming to construct an economic bus system, while the last criterion is proposed to make the bus system rationalized and humanized.

3.2 System Overview. In order to take the long and short distance of travel demand into consideration altogether and differentiate between buses and sharing-bikes, we propose our B&B system and give an overview of it in Figure 1. We divide our system into mainly two parts as follows:

Pre-Processing. In this part, we detect hot-spots from OD to determine bus station candidates and then build the flow network from the identified hot-spots. Each bus station candidate covers a clustering area and aggregates the flow within it.

Route Designing. Based on the flow network, we first initiate the starting road segment for each bus route. Then, we expand the bus route dynamically with the consideration of the sharing-bikes in each step. After that, we prune and supplement our system by adding more nearby station candidates, which can further promote the coverage rate of total flow.

4 System

4.1 Pre-processing. In the pre-processing, We conduct a two-step clustering to formulate the flow network from OD data. The two steps are designed as follow:

Step 1: Hot-spot Detection. There are two small steps in detection because DBSCAN [17] can only detect high-density clusters, and its performance depends on the parameter of *neighbor radius*(ϵ) and *minimum number of points in neighborhood*($minpts$), which is difficult to identify and make the results meet our needs. Firstly, the OD points are clustered by DBSCAN to detect the density centers, which are taken as hot-spots and bus station candidates. Then, clusters are exaggerated by aggregating nearby outliers to identify the coverage of each station candidates. The number of the bus station is usually 200-300, and the distance of each two bus sta-

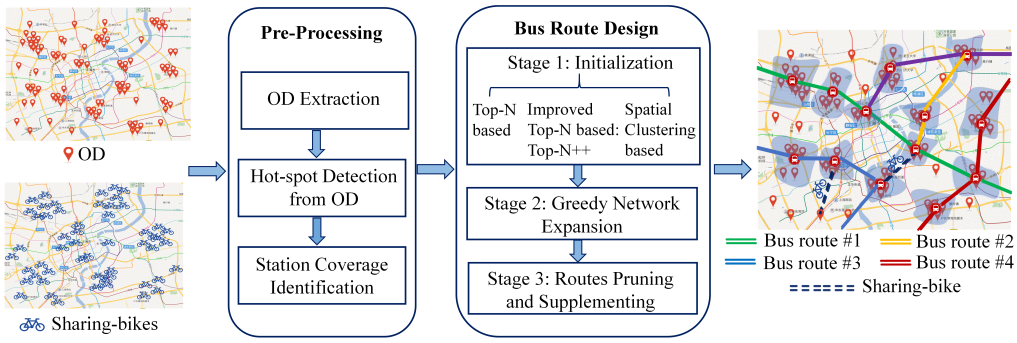


Figure 1: System overview

tion is usually 3km-5km in a modern city [1]. With this prior knowledge as a condition, we try different parameters to detect hot-spots as bus station candidates. The result parameters are $\epsilon = 100$m, $minpts$=110.

Step 2: Bus Station Identification: After deriving several hot-spots, we expand the coverage of each cluster by aggregating OD to their nearest cluster centers. With the identified clusters, we establish the flow network based on hot-spots. To be specific, for the distance matrix $\boldsymbol{D}$, the distance between the center of clusters represents for the distance between those hot-spots. For the passenger flow matrix $\boldsymbol{F}$, we match our extracted OD-pairs with points in our cluster to determine the passenger flow between hot-spots. It is worth noting that the graph is weighted and undirected, because both a trajectory from vertex i to j and from vertex j to i are aggregated to the flow between vertex i and j. It is reasonable to build such a undirected graph, as the bus route is bidirectional, which can serve the traffic demand from different directions.

4.2 Bus Route Design. The overall algorithm consists of 3 main components: Firstly, we initialize the routes with different ways of initialization with different intentions. Then, we propose a greedy network expansion algorithm considering sharing-bikes. Finally, we prune and supplement the existing bus routes as a balance between travel flow demand and efficiency.

4.2.1 Route Initialization. Given the number of the bus route N, we need to find N road segments with each one stands for a route as a start. In order to illustrate the point clearly, we propose three kinds of initialization approaches as follows:

Top-N Based Initialization. Since the flow coverage is our objective to optimize, Top-N initialization, which selects N routes with maximum passenger flow per unit distance denoted by $\rho_{ij} = \frac{\boldsymbol{F}_{ij}+\boldsymbol{F}_{ji}}{\boldsymbol{D}_{ij}}$, is a natural way for initialization. In this method, the road segments with the highest ρ_{ij} will be included into the route. Nevertheless, the road segments with the highest ρ_{ij} are often concentrated on particular areas, which will make the road segments likely to be connected, while reduces the search space and may miss some important areas [4].

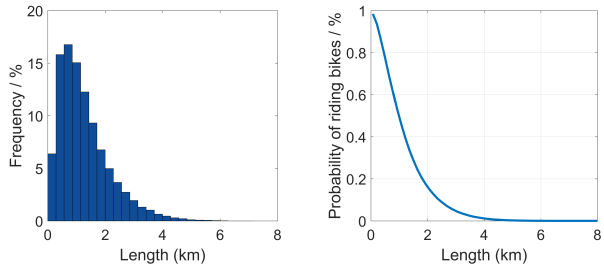
Agglomerative Clustering Based Initialization. In order to increase the coverage of the hot-spots, the agglomeration hierarchical clustering method [18] will be applied here to split the hot-spots into N tiny parts. Then we find the road segment with the maximum ρ_{ij} in each smaller segment. It has an advantage that the road segments belong to different part of the city, thus they are not connected and avert the deficiencies arise in Top-N initialization.

Top-N++ Initialization. When taken the passenger flow and the initial hot-spot coverage into account simultaneously, we ought to make a trade-off between the above two factors. To achieve this, we propose a novel way of initialization based on the Top-N initialization. We denote the distance between road segments $r_i = (v_1^i, v_2^i)$ and $r_j = (v_1^j, v_2^j)$ as follows,

$$dist(r_i, r_j) = \sum_{m,n\in\{1,2\}} dist(v_m^i, v_n^j)/4. \tag{4.4}$$

Similar to K-Means++ algorithm [3] which aims at scattering the initial centers of clusters, our Top-N++ approach is designed to disperse the initial road segments with the highest ρ. To achieve this, we form the candidate set for road segments, the size of which is k times to the number of routes N. Then, we use a probability-based method to pick N road segments from the set. The probability for picking a particular road segment is in proportion of its distance to the nearest road segment in the set. In this way, we select the initial N road segments and update the candidate set of road segments consecutively. Since random numbers are used for choosing road segments, we propose the initialization with multiple times and choose the set of N road segments with maximum average pairwise distance to ensure the dispersion of the road segments.

4.2.2 Greedy Network Expansion. In our algorithm, based on the route initialization, we propose a heuristic algorithm to add the relatively direct road seg-



(a) Bike Trip Length Distribution of New-York

(b) The Relation between the Prob. of Taking Bus and Trip Length

Figure 2: Mobility Pattern about Sharing-bike Users.

ments with higher flow gain on passenger flow per unit distance to the route network successively. Since we aim to consider the riding of sharing-bikes during bus route design process, the critical point is to determine when to select a road segment as a bus route and when passengers would ride bikes.

We first define some crucial notations: l_c is the length of the road segment c, i_c ($i_c \leq N$) is the ID of the route in which the road segment is added to, $p(l)$ represents for the relationship between the distance and the probability of riding sharing-bikes in a particular travel and E_c stands for the expected increase on number of people taking bus after adding the road segment c, Δf_c stands for the flow gain of adding a road segment c, Δg_c represents the score that measures the flow gain of adding a road segment c, $\Delta g'_c$ expresses for the score of flow gain w.r.t. bus passengers.

First we need to study the travel patterns on the trips of sharing-bikes. Figure 2(a) (a, b) shows the travel distances distribution $f(l)$ of sharing-bike [4]. Based on the statistical result of the sharing-bike trips, we define the relationship between the travel distance l and the probability of riding a bike p as follows,

$$p(l) = Pr(l' \geq l) = 1 - \int_0^l f(l')\ dl'. \tag{4.5}$$

The graph of function $p(l)$ is shown in Figure 2(b) (c). We find that less passengers choose to ride bikes when the trip distance increases. As there is an inverse relationship between the people taking buses and riding bikes, we use the probability function $p(d)$ to measure E_c as Equation 4.6.

$$E_c = \Delta f_c \cdot (1 - p(l_c)). \tag{4.6}$$

After considering the mobility pattern of the sharing-bike users as well as the relatively low speed of the bikes, there should be a maximum distance l_m for the trip of sharing-bikes. Also, when picking the optimal road segments, the segment's length and its relationship with the current route should not be overlooked when seeking for maximum gain. As for directness factor, the directness of the route L_{i_c} in which the road segment is added is calculated as $r(v_1^{i_c}, v_{s_{i_c}}^{i_c})$. Then, we can derive the expression for Δg_c as Equation 4.7,

$$\Delta g_c = \frac{\Delta f_c}{l_c \cdot r(v_1^{i_c}, v_{s_{i_c}}^{i_c})}. \tag{4.7}$$

When considering sharing-bikes, the score of gain w.r.t. bus passengers $\Delta g'_c$ is shown as Equation 4.8,

$$\Delta g'_c = \Delta g_c \cdot (1 - p(l_c)) = \frac{E_c}{l_c \cdot r(v_1^{i_c}, v_{s_{i_c}}^{i_c})}. \tag{4.8}$$

The detail for the algorithm is shown in Algorithm 1. In our network expansion algorithm, we first find all the road segments that serve as the extension of the current routesand satisfy the criteria in Equation 3.3. Next, we divide the road segments into two categories $\boldsymbol{l}, \boldsymbol{s}$ as the road segments in $\boldsymbol{s}$ are shorter than l_m and the segments in $\boldsymbol{l}$ are longer than l_m. For each category, we calculate the optimal road segments based on Equation 4.7 as c_l, c_s, the maximum score for flow gain Δg_c as $\Delta g_{max,s}, \Delta g_{max,l}$, the length of optimal road segments as l_{c_l}, l_{c_s} respectively.

When comparing the optimal result in two categories, if the Δg_c for c_s is greater than that of the long road segment c_l, but $\Delta g'_c$ for c_s is smaller, it implies that there is sufficient demand to construct the route, but few people are willing to take the bus. Under this circumstance, bike rides are utilized to provide transportation services in this road segment. Otherwise, bus routes are planned regularly, as we compare the value of $\Delta g_{max,s}, \Delta g_{max,l}$, and choose the corresponding segment with higher score.

4.2.3 Routes Pruning and Supplementing. In the route spanning method, since we pick the road segments with the higher gain on passenger flow per unit distance, then it may pass through some stations that are not so critical. To make further improvement in the coverage of hot-spots and passenger flows, we could add some passed stations to the whole system.

The rule of pruning and supplementing is shown in Algorithm 2. For those candidate stations which are not added to the bus routes currently but the constraints on both budget and directness would not be violated after adding to the route, we can include them to bus routes.

To better exemplify the role of pruning and supplementing, we choose a particular bus route to visualize the dynamic process in Figure 3. In Figure 3(a), the route is relatively straight but it fails to include passed stations. In Figure 3(b), in the red rectangle labeled area, passed stations are supplemented to satisfy more traffic flow, but the directness of the whole route slightly increases, which is the trade-off between meeting travel demand and improving efficiency.

As for sharing bikes, when the distance between a particular station and its nearest station in bus routes is within l_m, then the corresponding travel demand between it and stations in routes can be satisfied via

Algorithm 1 Network Expansion with Sharing-bikes

Input: Initialized bus routes $\boldsymbol{L} = \{L_1, L_2, ..., L_N\}$, Length Budget B_m, Directness Parameter α, Maximum distance l_m.
Output: Result bus routes $\boldsymbol{L}$, Sharing-bike rides $\boldsymbol{B}$.

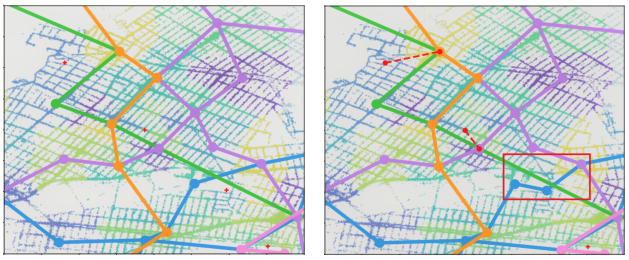
begin
- **Initialization:** Sharing-bike services $\boldsymbol{B} \leftarrow \emptyset$;
  The budget of each bus route $B_i \leftarrow B_m - \boldsymbol{D}_{v_1^i, v_2^i}$.
- **do**
  - **Calculate passenger flow** f based on Equation 3.3.
  - **Determine candidate set for road segments** C_r.
  - **Reset** $c_l, c_s \leftarrow \emptyset$, $\Delta g_{max,s}, \Delta g_{max,l} \leftarrow 0$, $l_{c_s}, l_{c_l} \leftarrow 0$.
  - **for** $c \in C_r$ **do**
    - Retrieve overall routes $\boldsymbol{R}$ based on $\boldsymbol{L} \cup \boldsymbol{B} \cup \{c\}$.
    - **Calculate the score** Δg_c based on Equation 4.7.
    - **if** $l_c < l_m$ & $\Delta g_c > \Delta g_{max,s}$ **then**
      - $\Delta g_{max,s} \leftarrow \Delta g_c$, $l_{c_s} \leftarrow l_c$, $c_s \leftarrow c$.
    - **else if** $\Delta g_c > \Delta g_{max,l}$ **then**
      - $\Delta g_{max,l} \leftarrow \Delta g_c$, $l_{c_l} \leftarrow l_c$, $c_l \leftarrow c$.
  - **if** $\Delta g_{max,l} > \Delta g_{max,s}$ **then**
    - $L_{c_i} \leftarrow L_{c_i} \cup \{c_l\}$, $B'_{c_i} \leftarrow B'_{c_i} - l_{c_l}$
  - **else**
    - **Calculate the** $\Delta g'_c$ **for both** c_s **and** c_l.
    - **if** $\Delta g_{c_s} > \Delta g'_{c_l}$ **then**
      - $L_{c_i} \leftarrow L_{c_i} \cup \{c_s\}$, $B'_{c_i} \leftarrow B'_{c_i} - l_{c_s}$.
    - **else**
      - $\boldsymbol{B} \leftarrow \boldsymbol{B} \cup \{c_s\}$// *Flow will be covered by bikes.*
- **While** $C_r \neq \emptyset$
- return $\boldsymbol{L}, \boldsymbol{B}$

riding sharing-bikes to nearest bus stations and taking buses then. In Figure 3(b), sharing-bike travels are shown with those red dashed lines, which can further elevate the coverage of flows and stations of our system.

Algorithm 2 Pruning and Supplementing for Bus Routes

Input: Bus routes $\boldsymbol{L} = \{L_1, L_2, ..., L_N\}$, Budget remaining of each bus route $B_{remain} = \{B_1, B_2, ..., B_N\}$, Length Budget B_m, Set of hot-spots $\boldsymbol{V} = \{v_1, v_2, ..., v_n\}$, Directness Parameter α.
Output: Pruned bus routes $\boldsymbol{L}$.

begin
- **Construct set of stations covered by bus routes** S_{cover};
- **for** $v \in \boldsymbol{V}$ **do**
  - **if** $v \notin S_{cover}$ **then**
    - **for** $L_i \in \boldsymbol{L}$ **do**
      - **Finding the nearest station to** v **in** L_i: $j \leftarrow \underset{v_i^j, v_i^{j+1} \in L_i}{\text{argmin}} [d(v_i^j, v) + d(v_i^{j+1}, v)]$;
      - **Calculate the increase of length**: $\Delta l \leftarrow \boldsymbol{D}_{v_i^j, v} + \boldsymbol{D}_{v_i^{j+1}, v} - \boldsymbol{D}_{v_i^j, v_i^{j+1}}$;
      - **if** $\Delta l > (\alpha - 1)\boldsymbol{D}_{v_i^j, v_i^{j+1}}$ & $\Delta l < B_i$ **then**
        - $L_i \leftarrow L_i \cup \{v\}$, $S_{cover} \leftarrow S_{cover} \cup \{v\}$; //*Adding the station to bus route*
        - $B_i \leftarrow B_i - \Delta l$;
        - **Break**
- return $\boldsymbol{L}$



(a) After expansion (b) After pruning and supplementing

Figure 3: An example of bus route design process.

5 Evaluation

In this section, we conduct extensive experiments on two different real-world datasets to answer the following research questions:

RQ1: Whether our proposed B&B system considering the sharing-bikes can improve the performance of serving travel demand with limited budgets?

RQ2: Whether our proposed route expansion algorithm is effective to increase the coverage of travel compared with other methods?

RQ3: How does the parameter, i.e., the length budget (B_m), the maximum number of bus routes (N), the upper bound of directness (α), influence the result of the planning routes?

RQ4: How does our proposed initialization algorithm perform under the different set of budgets?

5.1 Experimental Settings

5.1.1 Datasets. We utilize two kinds of dataset: *Green Taxi Trip Data* [2] and *Citi Bike Trip Data* [10, 23] from New York City, one of the largest city in the world, for evaluation. The key characteristics of these datasets are summarized in Table 2. The detail of two datasets with the pre-processing procedures are introduced as follows.

We filter the OD pairs with the travel time less than 1 minutes or the travel length less than 500m, which be a noise record caused by driver's mis-operation. After that, more than 1 million taxi OD records are obtained. Such two large-scale datasets in New York City guarantee the accuracy of our evaluation. To speed up the pre-processing steps, we divide the city into $10m \times 10m$ grids and only the geographical center of the grid is reserved if any OD locates within it. Then, we detect the hot-spots from grid OD and identify the bus station candidates via our proposed pre-processing method. Finally, we build the flow network for bus route expansion. As a result, there are 246 clusters in total.

5.1.2 Baselines. To evaluate the performance of our system, we compare our system B&B with several baselines including variants of our method without considering sharing-bikes. For the most state-of-the-art

solution, it proposed a system to build a loop line among OD, which is to route bi-directional bus routes [7]. All the baselines are described as follows.

L&NB: Planning bus routes by loop line method without considering the sharing-bike [7]. The loop lines are also common when designing bus routes.

L&B: Planning bus routes by loop line method [7] with the sharing-bikes.

B&NB: Planning bus routes by our system without the sharing-bikes.

B&B+EP: Planning bus routes by our system with the sharing-bikes only in expansion stage.

B&B+SP: Planning bus routes by our system with the sharing-bikes only in supplementing stage.

5.1.3 Parameters and Metrics. When studying the effect of different factors on our planning algorithm, the default parameters are set as Table 3. Besides, the default way of initialization is agglomerative clustering algorithm. To evaluate how our system satisfy the goal we aim to achieve, we use the following three metrics:

The Flow Coverage Rate (FCR): The ratio of the flow covered by our system to the total flow extracted from OD.

The Station Coverage Rate (SCR): The ratio of the number of bus station included by our routes to the number of all bus station candidates.

The Average Travel Time (ATT): The average travel time from all origin to destination flow satisfied by our bus routes. We fix the speed of the bus as well as bike by Table 3 to calculate the travel time.

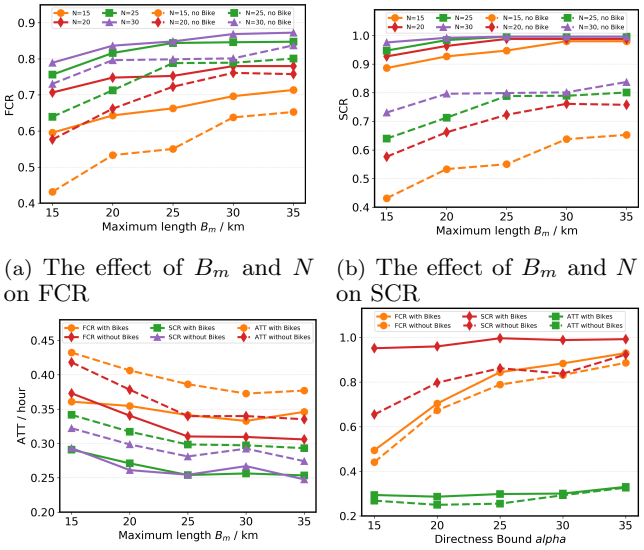
5.2 The Optimal Solution for Designing Bus Routes. For a large city like New York, to plan the bus routes is a complex task, as the planner should come out of a way to design bus routes efficiently and economically. In our case, we mainly tune two parameters: the maximum length B_m for routes and the number of routes N by grid search to find the optimal solution. We tune the maximum length B_m from 15km to 35km and tune the number of routes N from 15 to 30. In this case, directness bound α is set to 1.5.

Table 2: Summary of basic statistics of two representative datasets in New York City.

Service	Time	# of Records
Taxi Trip Data	Aug. 1st-31st, 2014	2,639,500
Bike Trip Data	Aug. 1st-16th, 2014	473,621

Table 3: The Default Parameters in our experiments.

Parameter	Value
Maximum trip length l_m of Sharing-bikes	$2km$
The average speed of buses v_b	40km/h
The average speed of Sharing-bikes v_{sb}	15km/h
The average speed of walking v_w	5km/h



(a) The effect of B_m and N on FCR

(b) The effect of B_m and N on SCR

(c) The effect of B_m and N on ATT

(d) The effect of α

Figure 4: The effect of parameters in our algorithm.

Figure 4 gives the result of the comparison with different budgets. We have the following conclusions:

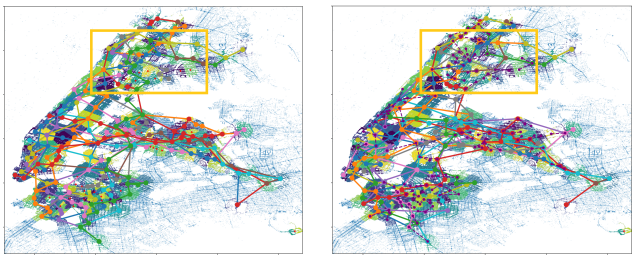
1) With the increase of the B_m and N, FCR and SCR both rise and ATT declines. However, the changes of the them slow down subsequently as B_m increases, which indicates that a relatively small budget would be enough to cover the traffic demand. The reason behind it is that the road segments with the highest flows are chosen first in our method. What's more, most of the stations will be included into the route when B_m are large, thus making the travel efficient enough and making it more difficult to add new stations into the route.

2) The performance of our methods with sharing-bikes is better than those without them, but the gap between the result has narrowed when B_m becomes larger. It is because that when the budget becomes larger, more bus stations will be covered. Therefore, the excessive travel demand completed by sharing-bike rides will be lesser in both the expansion stage and the supplementing stage.

From the discussion above, we find that B_m=25km and N=25 would be an appropriate setting for the planning of the bus routes in New York City. When N is fixed and B_m increases 40%(from 25km to 35km), the gain on FCR is merely 0.4%, on ATT is merely 0.2% and SCR just not change. Similarly, when B_m is fixed and N increases 20%(from 25 to 30), the gain on FCR and ATT is merely 0.5% and 0.2% and SCR still not change. Those are enough to demonstrate that this setting can achieve a trade-off between providing efficient and economic service for planners. Moreover, compared with the method without sharing-bikes, the result under this setting shows that FCR has been promoted by 7.05%, SCR has been improved by 15.55%

and ATT has been reduced by 17.32%, which justifies the effectiveness of the implementation of sharing-bikes.

To affirm the effect of sharing-bikes, we show the routes planning results for New York City with the derived optimal settings in Figure 5. Utilizing our proposed algorithm, the designed bus routes would cover most areas in the city to meet passengers' travel demand. However, it is worth noting that the result is more reasonable when taking sharing-bikes into consideration. Take the yellow rectangle labeled area as an example. Bus stations are unable to cover all stations in Figure 5(a). Actually, these short-distance routes should be improved, because sharing-bikes can better meet such travel demand as Figure 5(b) shown.

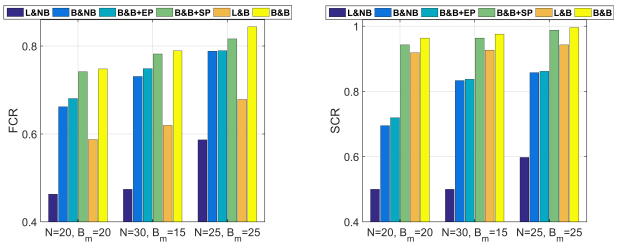


(a) Without sharing-bikes (b) With sharing-bikes

Figure 5: A case result for route planning for New York City, where dots represent bus stations, lines different colors represent different bus routes and the short purple lines represent that passengers would ride sharing-bikes to their nearest bus station.

5.3 The Benefits of Considering Sharing-bikes. We compare our system B&B with B&NB, B&B+EP and B&B+SP to further evaluate the benefits of considering sharing-bikes in different stages of our system. Results under different groups of parameters are shown Table 6 with $\alpha = 1.5$. From the results, we can observe that our B&B method always outperforms B&NB, B&B+EP, and B&B+SP. Specifically, FCR has been improved by about 2.20% (on average, the same below) and SCR has been improved by 1.20% about when sharing-bikes are considered in routes expansion, which means our decision function plays an important role in trading-off between bus and bike. Moreover, FCR has been improved by about 7.54% and SCR has been improved by 21.77% about when sharing-bikes are considered in routes supplementing, which means our designed bus can meet more traffic demand when considering that passengers would ride bikes to their bus station. By combining these two aspects together, our B&B method improves FCR by an SCR by 9.74% and 22.97%, which can serve more traffic flow under the same conditions.

5.4 The effect of expansion method. We also compare the results with and without considering sharing-bikes. The length limit of loop lines is relatively higher. In this case, we set the limit to $B_{m,loop} = 50km$.



(a) Comparison of FCR with different expansion methods (b) Comparison of SCR with different expansion methods

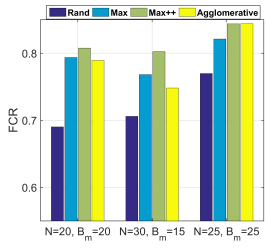
Figure 6: Comparison of different expansion methods.

The results are also shown in Figure 6, which verifies our algorithm. Obviously, our B&B method outperform others on both FCR and SCR with the different set of budget constraints. In addition, the increase of FCR exceeds 20% in almost all the cases regardless of sharing-bikes, which demonstrates that our expansion method can meet more traffic flow.

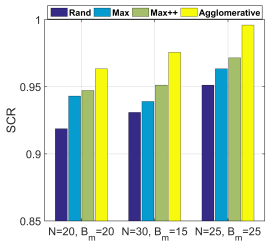
5.5 The Influence of Directness Parameter. The parameter α plays a crucial part in the design of the bus route. When α increases, the restriction on directness is relaxed and the number of candidate road segments in the network expansion increases, resulting in covering more stations. However, when α is higher, passengers will take a longer time from the source to destination, especially for long-distance travelers.

From the figure 4(d), we can observe that our FCR and SCR are always higher than that without bikes, while our ATT is always lower, which demonstrates that our system can increase the satisfaction for traffic flow and decrease the travel time at the same time. It is also worth noting that when α increases, the travel time first decreases than increases, which means that we can achieve a trade-off between providing efficient and satisfactory service for a passenger.

5.6 Comparisons of Initialization Methods. Table 7 further exhibits the quantitative results of the different initialization methods. We compare our result with that of random initialization, where we select N of road segments with the length less than $B_m/2$ randomly. To eliminate the variance, we run random method and Top-N++ method for 5 times and take the average of results. Also, we tune the size multiplier k for Top-N++ in [2, 3, 4, 5, 6] and report the best performance. From the result, it is obvious that all of three methods outperform the random one. The agglomerative clustering method, aiming at separating the initialized road segments, can enhance the coverage of stations as well as city areas, especially when the budget is limited. On the other hand, we find that Top-N++ achieves higher FCR in all of the conditions than Top-N. Therefore, when the flow coverage becomes the objective for optimization, Top-N++ can be utilized for better results.



(a) Comparison of FCR with different initializations



(b) Comparison of SCR with different initializations

Figure 7: Comparison of different initializations.

5.7 Implication and Application of our method.

In our method, we aim to design a new bus route system with the sharing-bikes for New York City. However, we do not take current bus routes into consideration. Therefore, to put our design into real practice, if the bus route is contained into current system, we suggest that it should be reserved. Meanwhile, if the designed bus route is not in the current system, we recommend that this route be replenished to current routes.

6 Conclusion

In this paper, we investigated the problem of bus routes planning with the sharing-bikes. Our goal is to maximize the flow coverage under three practical constraints: 1) the number of routes, 2) the length of each route, and 3) the directness of each road. Based on the flow extracted from large-scale OD data, we propose a heuristic approach to expand bus routes. Experiments on two large-scale datasets show the gains of our system and demonstrate that our system enables the city planners to design better public transportation systems.

References

[1] *Urban road traffic planning and design specification (1995–2015).* `http://kns.cnki.net/kns/detail/detail.aspx?FileName=SCSDGB50220-1995&DbName=SCSD`, 1995.

[2] *New york 2014 green taxi trip data.* `https://data.cityofnewyork.us/Transportation/2014-Green-Taxi-Trip-Data/2np7-5jsg`, 2018.

[3] D. Arthur and S. Vassilvitskii, *k-means++:the advantages of careful seeding*, in Eighteenth Acm-Siam Symposium on Discrete Algorithms, 2007.

[4] J. Bao, T. He, S. Ruan, Y. Li, and Y. Zheng, *Planning bike lanes based on sharing-bikes' trajectories*, in Proc. ACM SIGKDD, 2017, pp. 1377–1386.

[5] K. C. Brata, D. Liang, and S. H. Pramono, *Location-based augmented reality information for bus route planning system*, International Journal of Electrical and Computer Engineering, 5 (2015), pp. 142–149.

[6] C. Chen, D. Zhang, N. Li, and Z. H. Zhou, *B-planner: Planning bidirectional night bus routes using large-scale taxi gps traces*, IEEE Transactions on Intelligent Transportation Systems, 15 (2014).

[7] S. P. Chuah, H. Wu, Y. Lu, L. Yu, and S. Bressan, *Bus routes design and optimization via taxi data analytics*, in Proc. ACM CIKM, 2016, pp. 2417–2420.

[8] R. Z. Farahani, E. Miandoabchi, W. Szeto, and H. Rashidi, *A review of urban transportation network design problems*, European Journal of Operational Research, 229 (2013), pp. 281 – 302.

[9] Y. Ge, H. Xiong, A. Tuzhilin, K. Xiao, M. Gruteser, and M. Pazzani, *An energy-efficient mobile recommender system*, in Proc. ACM SIGKDD.

[10] Y. Li, Y. Zheng, H. Zhang, and L. Chen, *Traffic prediction in a bike-sharing system*, in Proc. ACM SIGSPATIAL, 2015.

[11] W. Liu, Y. Zheng, S. Chawla, J. Yuan, and X. Xie, *Discovering spatio-temporal causal interactions in traffic data streams*, in Proc. ACM SIGKDD, 2011.

[12] Y. Liu, C. Liu, N. J. Yuan, L. Duan, Y. Fu, H. Xiong, S. Xu, and J. Wu, *Exploiting heterogeneous human mobility patterns for intelligent bus routing*, in Proc. IEEE ICDM, 2015.

[13] S. Ma, Y. Zheng, and O. Wolfson, *T-share: A large-scale dynamic taxi ridesharing service*, in Proc. ICDE, 2013.

[14] R. Nair, E. Miller-Hooks, R. C. Hampshire, and A. Bui, *Large-scale vehicle sharing systems: Analysis of vlib'*, International Journal of Sustainable Transportation, 7 (2013).

[15] C. Partha, *Genetic algorithms for optimal urban transit network design*, Computer-Aided Civil and Infrastructure Engineering, 18.

[16] S. Rinzivillo, S. Mainardi, F. Pezzoni, M. Coscia, D. Pedreschi, and F. Giannotti, *Discovering the geographical borders of human mobility*, KI - Künstliche Intelligenz, 26 (2012).

[17] P. N. Tan, M. Steinbach, and Vipin, *Introduction to data mining*, Posts Telecom Press, 2006.

[18] J. Wardjr, *Hierarchical grouping to optimize an objective function*, Publications of the American Statistical Association, 58 (1963), pp. 236–244.

[19] F. Wu and Y. Xue, *Innovations of bike sharing industry in china : A case study of mobikes stationless bike sharing system*, 2017.

[20] H. Yao, F. Wu, J. Ke, X. Tang, Y. Jia, S. Lu, P. Gong, J. Ye, and Z. Li, *Deep multi-view spatial-temporal network for taxi demand prediction*, arXiv preprint arXiv:1802.08714, (2018).

[21] J. Yuan, Y. Zheng, and X. Xie, *Discovering regions of different functions in a city using human mobility and pois*, Proc. ACM SIGKDD, (2012).

[22] F. Zhao and X. Zeng, *Simulated annealing-genetic algorithm for transit network optimization.*, Journal of Computing in Civil Engineering, 20 (2006).

[23] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, *Urban computing: Concepts, methodologies, and applications*, ACM Trans. Intell. Syst. Technol., 5 (2014).

[24] Y. Zheng, Y. Liu, J. Yuan, and X. Xie, *Urban computing with taxicabs*, in Proc. Ubicomp, ACM, 2011.